

## NOTES AND COMMENTS

**On the surjectivity of the mapping between utilities and choice probabilities**

ANDRIY NORETS

Department of Economics, Princeton University

SATORU TAKAHASHI

Department of Economics, National University of Singapore

This note considers a standard multinomial choice model. It is shown that if the distribution of additive utility shocks has a density, then the mapping from deterministic components of utilities to choice probabilities is surjective. In other words, any vector of choice probabilities can be obtained by selecting suitable utilities for alternatives. This result has implications for at least three areas of interest to econometricians: the Hotz and Miller (1993) estimator for structural dynamic discrete choice models, nonparametric identification of multinomial choice models, and consistency of conditional density estimators based on covariate dependent mixtures.

KEYWORDS. Multinomial choice models, identification, Hotz and Miller estimator, covariate dependent mixtures.

JEL CLASSIFICATION. C35, C61.

## 1. MAIN RESULT AND ITS IMPLICATIONS

Multinomial choice models were introduced in economics by McFadden (1974) and since then have been extensively used in applications. In the model, the agent chooses between  $J + 1$  possible alternatives denoted by  $\{0, 1, \dots, J\}$ . The utility from choosing alternative  $j$  is given by  $u_j + \varepsilon_j$ , where  $u_j$  is a deterministic component of the utility, which might depend on agent's and alternative's characteristics, and  $\varepsilon_j$  is a random shock. Let us denote the distribution of  $\varepsilon = (\varepsilon_0, \dots, \varepsilon_J)$  by  $G$  and assume that it satisfies the following assumption.

ASSUMPTION 1.  *$G$  is absolutely continuous with respect to the Lebesgue measure, that is,  $G$  has a density with respect to the Lebesgue measure.*

Multinomial logit and probit models are most commonly used in applications. The former is obtained when  $\varepsilon_j$ 's are independently identically distributed (i.i.d.) according

---

Andriy Norets: [anorets@princeton.edu](mailto:anorets@princeton.edu)

Satoru Takahashi: [ecsst@nus.edu.sg](mailto:ecsst@nus.edu.sg)

We are grateful to Bo Honore for helpful discussions. We thank anonymous referees for comments that helped to improve the manuscript.

Copyright © 2013 Andriy Norets and Satoru Takahashi. Licensed under the [Creative Commons Attribution-NonCommercial License 3.0](http://creativecommons.org/licenses/by-nc/3.0/). Available at <http://www.qeconomics.org>.

DOI: 10.3982/QE252

to an extreme value type I distribution; the latter is obtained when  $\varepsilon_j$ 's have a joint normal distribution.

Conditional on  $(u_0, \dots, u_J)$ , the probability of choosing alternative  $j$  is given by

$$p_j = \Pr(\varepsilon : u_j + \varepsilon_j \geq u_i + \varepsilon_i, \forall i \neq j). \quad (1)$$

A location normalization on  $u_j$ 's can be introduced without a loss of generality. Thus, let  $u_0 = 0$  and  $u = (u_1, \dots, u_J)$ . Equation (1) defines a mapping  $\phi : \mathbb{R}^J \rightarrow \Delta^J$ , where  $\Delta^J = \{p = (p_0, p_1, \dots, p_J) : p_j \geq 0, \sum_j p_j = 1\}$  is a  $J$ -dimensional simplex. (Ties,  $u_i + \varepsilon_i = u_j + \varepsilon_j$ ,  $i \neq j$ , occur with probability 0 due to Assumption 1.)

**THEOREM 1.** *Under Assumption 1, the image of  $\phi$ ,  $\phi(\mathbb{R}^J)$ , includes the interior of  $\Delta^J$ . If the support of  $G$  is bounded, then  $\phi(\mathbb{R}^J) = \Delta^J$ ; if the support is equal to  $\mathbb{R}^{J+1}$ , then  $\phi(\mathbb{R}^J)$  is equal to the interior of  $\Delta^J$ .*

**COROLLARY 1.** *Suppose  $G$  satisfies Assumption 1 and the support of  $G$  is connected.<sup>1</sup> Then  $\phi^{-1}$  is well defined on the interior of  $\Delta^J$ . If, in addition, the support of  $G$  is equal to  $\mathbb{R}^{J+1}$ , then  $\phi$  is a bijection between  $\mathbb{R}^J$  and the interior of  $\Delta^J$ .*

Hofbauer and Sandholm (2002) showed the surjectivity of  $\phi$  in their proof of Theorem 2.1 along with other results under the assumption that the density of  $G$  is strictly positive and  $\phi$  is continuously differentiable. In the following section, we present a proof of Theorem 1 under weaker Assumption 1, which does not require the full support for  $G$  and continuous differentiability of  $\phi$ . Beyond this technical improvement, the purpose of this note is to bring the surjectivity result to econometrics literature. Theorem 1 does not appear to be known in econometrics literature although it has important implications for at least three areas of interest to econometricians: the Hotz and Miller (1993) estimator for structural dynamic discrete choice models, nonparametric identification of multinomial choice models, and consistency of conditional density estimators based on covariate dependent mixtures. We briefly review these implications below.

A structural dynamic discrete choice model is a dynamic optimization model with discrete controls. Eckstein and Wolpin (1989), Rust (1994), Miller (1997), Aguirregabiria and Mira (2010), and Keane, Todd, and Wolpin (2011) surveyed the literature on applications of these models in empirical work. Rust (1987) introduced a specification of dynamic discrete choice model that directly extends (1) so that  $u_j$ 's are the deterministic components of the alternative specific value functions. Hotz and Miller (1993) proposed a computationally attractive estimation procedure for this specification, which together with its extensions (Aguirregabiria and Mira (2007b), Pesendorfer and Schmidt-Dengler (2008), and Bajari, Benkard, and Levin (2007)) is widely applied, especially in empirical industrial organization. Hotz and Miller (1993) showed in their Proposition 1 that if  $G$  has a positive continuous density on  $\mathbb{R}^{J+1}$ , then  $\phi$  is differentiable and has an inverse,  $\phi^{-1}$ , defined on its image,  $\phi(\mathbb{R}^J)$  (they did not show that the image,  $\phi(\mathbb{R}^J)$ , covers the

<sup>1</sup>The support of  $G$  is the set of  $\varepsilon \in \mathbb{R}^{J+1}$  such that  $G[O] > 0$  for any open neighborhood  $O$  of  $\varepsilon$ . Equivalently, the support of  $G$  is the smallest closed set with probability 1.

interior of the simplex; see line 7 in their proof of Proposition 1 for a precise statement of their result). Based on this result, [Hotz and Miller \(1993\)](#) proposed a two stage estimator. In the first stage of the procedure, the choice probabilities are estimated nonparametrically as functions of covariates (observed state variables). Let us denote this estimator by  $\hat{p}$ . In the second stage,  $\phi$  is inverted to obtain  $\hat{u} = \phi^{-1}(\hat{p})$ , which is then used to estimate structural parameters without ever solving the structural dynamic optimization model (solution of the optimization problem is very computationally intensive). [Hotz and Miller's \(1993\)](#) procedure is usually used with the assumption that  $\varepsilon_j$ 's are extreme value i.i.d. In this case,  $\phi$  and  $\phi^{-1}$  have known analytical expressions (multinomial logit; see [Domencich and McFadden \(1975\)](#) for derivations) and  $\phi$  is clearly surjective (its image is equal to the interior of  $\Delta^J$ ). Suppose an econometrician is not willing to use an extreme value distribution for  $G$  and uses another distribution instead, for example, a normal distribution. If the implied  $\phi$  were not surjective, then it could have happened that  $\phi^{-1}$  were not defined at  $\hat{p}$  and [Hotz and Miller's \(1993\)](#) procedure would be problematic. [Theorem 1](#) shows that this does not happen under [Hotz and Miller's \(1993\)](#) assumptions.

Next, let us consider implications of the theorem for identification of multinomial choice models. Sufficient conditions for semi- and nonparametric identification of  $u$  and  $G$  can be found in [Matzkin \(1991, 1993\)](#). [Theorem 1](#) shows that without additional restrictions on  $u$  as functions of covariates, the knowledge of  $p$  ( $p$  can be estimated from data on individual choices) does not imply any restrictions on  $G$ , at least in the class of distributions satisfying [Assumption 1](#). Thus, even partial (or set) identification of  $G$  is not possible without additional restrictions on  $u$ .

Finally, let us consider implications of [Theorem 1](#) for models based on covariate dependent mixture models. Covariate dependent mixtures, also known as mixtures of experts in statistics, were considered by [Jacobs, Jordan, Nowlan, and Hinton \(1991\)](#), [Jordan and Xu \(1995\)](#), [Peng, Jacobs, and Tanner \(1996\)](#), [Wood, Jiang, and Tanner \(2002\)](#), [Geweke and Keane \(2007\)](#), and [Villani, Kohn, and Giordani \(2009\)](#), among others. Consider the finite location–scale mixture model for a conditional density of  $y$  given covariates  $x$ ,

$$p(y|x, \mu, \sigma) = \sum_{j=0}^J \pi_j(x) f\left(\frac{y - \mu_j}{\sigma}\right), \quad (2)$$

where  $f$  is a density and  $(\pi_0(x), \dots, \pi_J(x))$  are mixing probabilities. [Norets \(2010\)](#) and [Norets and Pelenis \(2011\)](#) showed that if the mixing probabilities are flexible functions of covariates  $x$ , then the model in (2) can approximate and consistently estimate large nonparametric classes of conditional densities. As shown in [Geweke and Keane \(2007\)](#), a computationally convenient way to implement (2) is to use a multinomial choice model for  $(\pi_0(x), \dots, \pi_J(x))$  combined with polynomials for  $u(x) = (u_1(x), \dots, u_J(x))$ . Alternative flexible specifications for  $u(x)$  include splines and series expansions. If  $u(x)$  are modeled flexibly, then the [Hotz and Miller \(1993\)](#) results on the existence of the inverse of  $\phi$  and [Theorem 1](#) imply that the resulting mixing probabilities are flexible functions of covariates. [Corollary 3.1](#) of [Norets \(2010\)](#) gives a rigorous proof of an approximation result for (2) used with a multinomial logit model for  $(\pi_0(x), \dots, \pi_J(x))$ . [Theorem 1](#) implies that the result holds for any  $G$  satisfying the assumptions of [Hotz and Miller \(1993\)](#).

## 2. PROOF OF THEOREM 1

First, let us redefine the mapping from utilities to choice probabilities so that it has the domain equal to a  $J$ -dimensional simplex. This is achieved by using  $\log(x_j)$  instead of  $u_j$  inside the mapping from utilities to choice probabilities. Since  $u_j$ 's can have a location normalization, it is without loss of generality to restrict  $x_j$ 's to the simplex. More formally, for any finite  $u_j, j = 0, 1, \dots, J$ , let  $x_j = \exp(u_j - \log[\sum_{i=0}^J \exp(u_i)])$ . Then it is easy to see that (i) using  $\log(x_j), j = 0, \dots, J$ , as utilities results in the same choice probabilities as  $u_j$ 's, (ii)  $x_j > 0$ , and (iii)  $\sum_{j=0}^J x_j = 1$ . Let  $\text{int } \Delta^J = \{x = (x_0, x_1, \dots, x_J) : x_j > 0, \sum_{j=0}^J x_j = 1\}$  denote the interior of  $\Delta^J$  and let  $\text{bd } \Delta^J = \Delta^J \setminus (\text{int } \Delta^J)$  denote the boundary of  $\Delta^J$ . For each  $x = (x_0, x_1, \dots, x_J) \in \Delta^J$ , define  $\psi : \Delta^J \rightarrow \Delta^J$  by

$$\psi_j(x) = \begin{cases} \Pr(\varepsilon : \ln x_j + \varepsilon_j \geq \ln x_i + \varepsilon_i, \forall i \neq j), & \text{if } x_j > 0, \\ 0, & \text{if } x_j = 0. \end{cases}$$

It follows from this definition that  $\phi(\mathbb{R}^J) = \psi(\text{int } \Delta^J)$ . Also, for any  $x \in \text{bd } \Delta^J$ , we have  $\psi(x) \in \text{bd } \Delta^J$ . Thus, to prove Theorem 1, it is enough to show  $\psi(\Delta^J) = \Delta^J$ .

Note that by Assumption 1,  $\psi$  is continuous.<sup>2</sup> Also note that the restriction of  $\psi$  to  $\text{bd } \Delta^J$  does not map antipodal points to the same point. More precisely, for any  $x \in \text{bd } \Delta^J$ , denote by  $x^*$  the point in  $\text{bd } \Delta^J$  that is hit by the ray from  $x$  to  $(1/(J+1), \dots, 1/(J+1))$ . Then, for any  $x \in \text{bd } \Delta^J$ , we have  $\psi(x) \neq \psi(x^*)$  because  $\psi_j(x) = 0 < \psi_j(x^*)$  whenever  $x_j = 0$ . This property turns out to be sufficient to show  $\psi(\Delta^J) = \Delta^J$ . The proof is a standard application of the Borsuk–Ulam theorem in algebraic topology. For a textbook treatment of the theorem and related definitions and results, see, for example, [Fulton \(1995\)](#). Also, [Kojima and Takahashi \(2008\)](#) used a similar proof technique in their Lemma 1(d) to show the surjectivity of  $\phi$  when  $\varepsilon$  is independently distributed according to an exponential distribution.

Suppose, on the contrary, that there exists  $x^0 \in \Delta^J$  such that  $x^0 \notin \psi(\Delta^J)$ . For each  $x \in \Delta^J$ , define  $g(x)$  by the point in  $\text{bd } \Delta^J$  that is hit by the ray from  $x^0$  to  $\psi(x)$ . Then  $g : \Delta^J \rightarrow \text{bd } \Delta^J$  is a continuous function that coincides with  $\psi$  on  $\text{bd } \Delta^J$ . Thus the restriction of  $\psi$  to  $\text{bd } \Delta^J$  has degree 0.

Let  $F$  be a homeomorphism from  $\text{bd } \Delta^J$  to a  $(J-1)$ -dimensional sphere  $S^{J-1} = \{y \in \mathbb{R}^J : |y| = 1\}$  that preserves antipodal points, that is,  $F(x^*) = -F(x)$  for each  $x \in \text{bd } \Delta^J$ . For example,

$$F(x) = \frac{(x_1, \dots, x_J) - (1/(J+1), \dots, 1/(J+1))}{|(x_1, \dots, x_J) - (1/(J+1), \dots, 1/(J+1))|}.$$

Then  $h_0 = F \circ \psi \circ F^{-1} : S^{J-1} \rightarrow S^{J-1}$  is continuous, has degree 0, and satisfies  $h_0(y) \neq h_0(-y)$  for any  $y \in S^{J-1}$ .

Define a homotopy  $H : S^{J-1} \times [0, 1] \rightarrow S^{J-1}$ :

$$H(y, t) = \frac{h_0(y) - th_0(-y)}{|h_0(y) - th_0(-y)|}.$$

<sup>2</sup>Assumption 1 is only used to obtain continuity of  $\psi$  and zero probability of  $u_i + \varepsilon_i = u_j + \varepsilon_j, i \neq j$ . These properties can be obtained even under weaker assumptions, for example,  $\varepsilon_j$  are independent and have a Cantor distribution. Thus, *absolute* continuity of  $G$  is not required for the results.

Note that  $H$  is well defined and continuous since  $h_0(y) \neq h_0(-y)$ . We have  $h_0 = H(\cdot, 0)$ . Since  $h_1 := H(\cdot, 1)$  satisfies  $h_1(y) = -h_1(-y)$  for all  $x \in S^{J-1}$ , it follows from the Borsuk–Ulam theorem that  $h_1$  has an odd degree. This contradicts that  $h_0$  and  $h_1$  are homotopic.

### 3. PROOF OF COROLLARY 1

Mapping  $\phi$  was originally defined on  $\mathbb{R}^J$  under the location normalization for utilities. We define  $\tilde{\phi}(u) = \phi((u_1 - u_0, \dots, u_J - u_0))$  for  $u \in \mathbb{R}^{J+1}$ . Suppose for  $u, u' \in \mathbb{R}^{J+1}$ ,  $\tilde{\phi}(u) = \tilde{\phi}(u') = p \in \text{int } \Delta^J$ . To prove the claim of the corollary, it suffices to show that  $u_j - u_0 = u'_j - u'_0$  for any  $j$ .

Let  $J^* = \arg \max_{j=0, \dots, J} [u_j - u'_j]$ . If  $J^* = \{0, 1, \dots, J\}$ , then  $u_j - u_0 = u'_j - u'_0$  for any  $j$  and the corollary is proved. To obtain a contradiction, we assume  $\{0, \dots, J\} \setminus J^* \neq \emptyset$ .

Let  $O = \{\varepsilon \in \mathbb{R}^{J+1} : \arg \max_j [u_j + \varepsilon_j] \subset J^*\}$ . Similarly, let  $O' = \{\varepsilon \in \mathbb{R}^{J+1} : \arg \max_j [u'_j + \varepsilon_j] \cap J^* = \emptyset\}$ . Let us establish the following three properties of  $O$  and  $O'$ .

(i)  $O$  and  $O'$  are open in  $\mathbb{R}^{J+1}$ . For any  $\varepsilon \in O$ , any  $j \in J^*$ , and any  $k \notin J^*$ ,  $u_j + \varepsilon_j > u_k + \varepsilon_k$ , and if we change  $\varepsilon$  by a sufficiently small amount, the resulting optimal choice(s) conditional on  $u$  will still be in  $J^*$ . Thus,  $O$  is open. Similarly,  $O'$  is open.

(ii) Both  $O$  and  $O'$  have nonempty intersections with the support of  $G$ , denoted by  $S$ . If  $O \cap S = \emptyset$ , then  $p_j = 0$  for any  $j \in J^*$ , which contradicts  $p \in \text{int } \Delta^J$ . If  $O' \cap S = \emptyset$ , then  $p_j = 0$  for any  $j \notin J^*$ , which contradicts  $p \in \text{int } \Delta^J$ .

(iii)  $O \cup O' = \mathbb{R}^{J+1}$ . Suppose  $\varepsilon \notin O$ . Then some  $k \notin J^*$  is an optimal choice conditional on  $u$  and for any  $i \in \{0, \dots, J\}$ ,  $u_k - u_i \geq \varepsilon_i - \varepsilon_k$ . By definition of  $J^*$ ,  $u_k - u_j < u'_k - u'_j$  for any  $j \in J^*$ . Thus, for any  $j \in J^*$ ,  $u'_k - u'_j > \varepsilon_j - \varepsilon_k$  and  $j$  cannot be an optimal choice conditional on  $u'$ . Thus,  $\varepsilon \in O'$  and  $O \cup O' = \mathbb{R}^{J+1}$ .

By the definition of connectedness, there is no pair of sets  $O, O'$  such that (i)  $O, O'$  are both open in  $\mathbb{R}^{J+1}$ , (ii)  $O \cap S$  and  $O' \cap S$  are both nonempty, (iii)  $O \cup O' \supseteq S$ , and (iv)  $O \cap O' \cap S$  is empty. Since (i)–(iii) are true, (iv) must be false.

As  $O \cap O' \cap S \neq \emptyset$ , by the definition of the support, we have  $G[O \cap O'] > 0$ . Thus,  $\sum_{j \in J^*} \tilde{\phi}_j(u) = G(O) > G(\mathbb{R}^{J+1} \setminus O') = \sum_{j \in J^*} \tilde{\phi}_j(u')$ , which is a contradiction to  $\tilde{\phi}(u) = \tilde{\phi}(u')$ .

### REFERENCES

- Aguirregabiria, V. and P. Mira (2010), “Dynamic discrete choice structural models: A survey.” *Journal of Econometrics*, 156, 38–67. [150]
- Aguirregabiria, V. and P. Mira (2007b), “Sequential estimation of dynamic discrete games.” *Econometrica*, 75, 1–53. [150]
- Bajari, P., L. Benkard, and J. Levin (2007), “Estimating dynamic models of imperfect competition.” *Econometrica*, 75, 1331–1370. [150]
- Domencich, T. and D. McFadden (1975), *Urban Travel Demand: A Behavioral Analysis: A Charles River Associates Research Study*. Contributions to Economic Analysis. Elsevier, Amsterdam, Netherlands. [151]

Eckstein, Z. and K. Wolpin (1989), "The specification and estimation of dynamic stochastic discrete choice models: A survey." *Journal of Human Resources*, 24, 562–598. [150]

Fulton, W. (1995), *Algebraic Topology*. Springer, New York. [152]

Geweke, J. and M. Keane (2007), "Smoothly mixing regressions." *Journal of Econometrics*, 138, 252–290. [151]

Hofbauer, J. and W. H. Sandholm (2002), "On the global convergence of stochastic fictitious play." *Econometrica*, 70, 2265–2294. [150]

Hotz, J. and R. Miller (1993), "Conditional choice probabilities and the estimation of dynamic models." *Review of Economic Studies*, 60, 497–529. [149, 150, 151]

Jacobs, R. A., M. I. Jordan, S. J. Nowlan, and G. E. Hinton (1991), "Adaptive mixtures of local experts." *Neural Computation*, 3, 79–87. [151]

Jordan, M. and L. Xu (1995), "Convergence results for the EM approach to mixtures of experts architectures." *Neural Networks*, 8, 1409–1431. [151]

Keane, M., P. Todd, and K. Wolpin (2011), "The structural estimation of behavioral models: Discrete choice dynamic programming methods and applications." In *Handbook of Labor Economics*, Vol. 4, Chap. 4, 331–461, Elsevier, Amsterdam, Netherlands. [150]

Kojima, F. and S. Takahashi (2008), "p-Dominance and perfect foresight dynamics." *Journal of Economic Behavior and Organization*, 67, 689–701. [152]

Matzkin, R. (1993), "Nonparametric identification and estimation of polychotomous choice models." *Journal of Econometrics*, 58, 137–168. [151]

Matzkin, R. L. (1991), "Semiparametric estimation of monotone and concave utility functions for polychotomous choice models." *Econometrica*, 59, 1315–1327. [151]

McFadden, D. (1974), "Conditional logit analysis of qualitative choice behavior." In *Frontiers in Econometrics* (P. Zarembka, ed.), 105–142, Academic Press, New York. [149]

Miller, R. A. (1997), "Estimating models of dynamic optimization with microeconomic data." In *Handbook of Applied Econometrics*, Vol. 2 (M. Pesaran and P. Schmidt, eds.), Chap. 6, 246–299, Blackwell, Oxford, U.K. [150]

Norets, A. (2010), "Approximation of conditional densities by smooth mixtures of regressions." *The Annals of Statistics*, 38, 1733–1766. [151]

Norets, A. and J. Pelenis (2011), "Posterior consistency in conditional density estimation by covariate dependent mixtures." Unpublished manuscript, Princeton University. [151]

Peng, F., R. A. Jacobs, and M. A. Tanner (1996), "Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts models with an application to speech recognition." *Journal of the American Statistical Association*, 91, 953–960. [151]

Pesendorfer, M. and P. Schmidt-Dengler (2008), "Asymptotic least squares estimators for dynamic games." *Review of Economic Studies*, 75, 901–928. [150]

Rust, J. (1987), “Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher.” *Econometrica*, 55, 999–1033. [150]

Rust, J. (1994), “Structural estimation of Markov decision processes.” In *Handbook of Econometrics* (P. Engle and D. McFadden, eds.), North-Holland, Amsterdam, Netherlands. [150]

Villani, M., R. Kohn, and P. Giordani (2009), “Regression density estimation using smooth adaptive Gaussian mixtures.” *Journal of Econometrics*, 153, 155–173. [151]

Wood, S., W. Jiang, and M. Tanner (2002), “Bayesian mixture of splines for spatially adaptive nonparametric regression.” *Biometrika*, 89, 513–528. [151]

---

Submitted October, 2012. Final version accepted October, 2012.